AUSTRALIAN OPEN TENNIS CHAMPIONSHIP ANALYSIS

Table of Contents

Formats and Values of the Dataset

Formats/Attributes	Description
Year	Numeric, ordinal values representing the year of the championship matches
Gender	Categorical, nominal values indicating whether match was Women's or Men's
Champion	Categorical, nominal values containing the names of the champions
Champion Nationality	Categorical, nominal values representing the nationalities of the champions
Champion Country	Categorical, nominal values indicating the countries represented by the champions
Score	Categorical, ordinal values representing the match scores
Win-Rate	Numeric, ratio values representing the win rates of champions as percentages
1st Won , 1st Lost, 2nd Won, 2nd Lost, 3rd Won, 3rd Lost, 4th Won, 4th Lost, 5th Won, 5th Lost	Numeric, ratio values showing the win-loss records of champions in earlier rounds leading up to the final
Wins	Numeric, ratio values indicating the total number of matches won by the champions
Loss	Numeric, ratio values indicating the total number of matches loss by the champions
Runner-Up	Categorical, nominal values containing the names of the runner-up players (the opponents of the champions)
Runner-Up Nationality	Categorical, nominal values representing the nationalities of runner-up players
Runner-Up Country	Categorical, nominal values indicating the countries represented by the runner-up players

Characteristics of the Dataset

- The dataset covers championship (finals) matches of Australian Open Tennis championships from 1905 to 2023, providing a total of 118 years of data.
- Includes information about both Women's and Men's matches, with a detailed information about the champion's and runner-up's name, nationality and representing country as well as their respective scores throughout the match.
- Total wins and losses of champions is calculated after each match and is utilised to showcase its win rate.
- Win rates vary by year and gender, with some champions having higher win rates than others.
- Dataset covers scores of champions in each round with scores of 1st win, 1st loss, 2nd win, 2nd loss, 3rd win, 3rd loss for Women's match and additional 2 rounds with 4th win, 4th loss and 5th win, 5th loss for Male's match.
- Some champions have multiple championships over the years (e.g. Novak Djokovic).
- Win rate ranges between 48.39% to 88.89%
- Some champion had null values due to walkover.
- There are 7 champions who have won 5 games or more.

TRANSFORMATION/CALCULATION ON THE DATASET

To explore more insights and trends with the dataset, two calculation was performed. Firstly, Overall win-rate for each champion is calculated by dividing the total number of wins to the total number of games played. This calculation was implemented on both the dataset consisting of all champions and dataset consisting of the top players with more than 5 wins as an effective way to compare their performance. Secondly, win-rate for each individual set of matches is calculated by dividing the number of wins in the set with number of games played in the set. With such, outliers and trend can be easily seen and visualized using the graphs presented within this report.

NATIONALITIES TREND

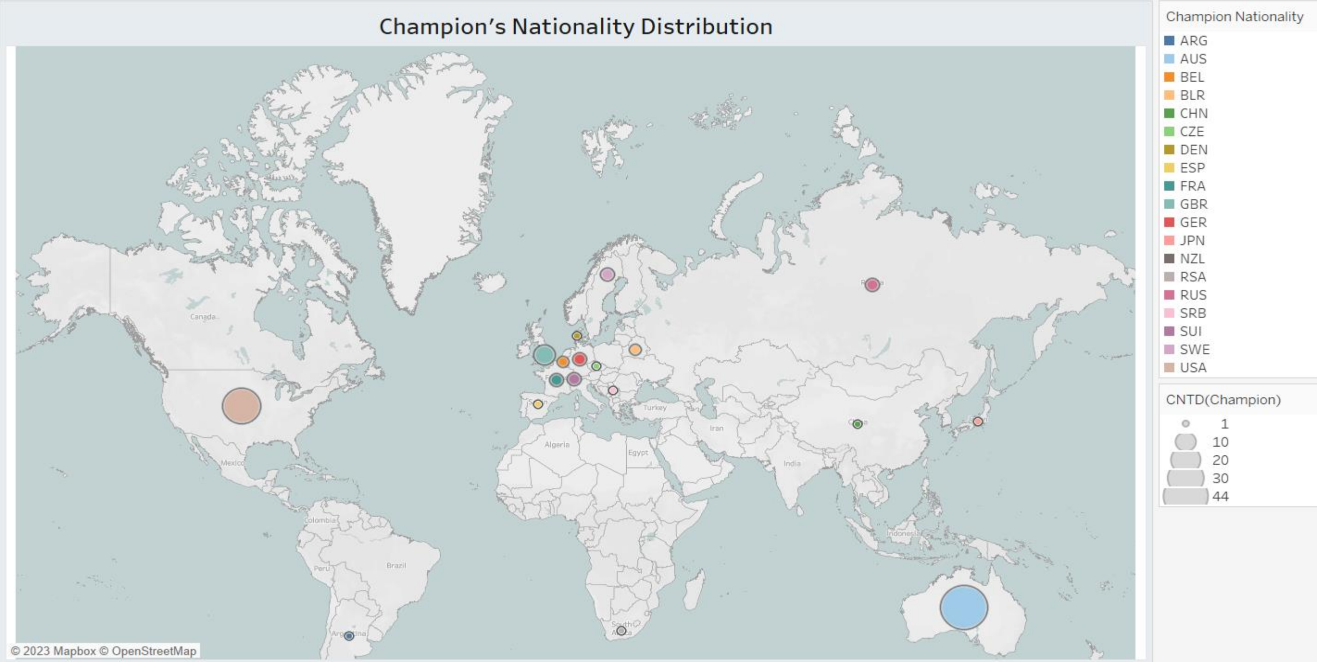

Figure 1: Champion's Nationality Distribution

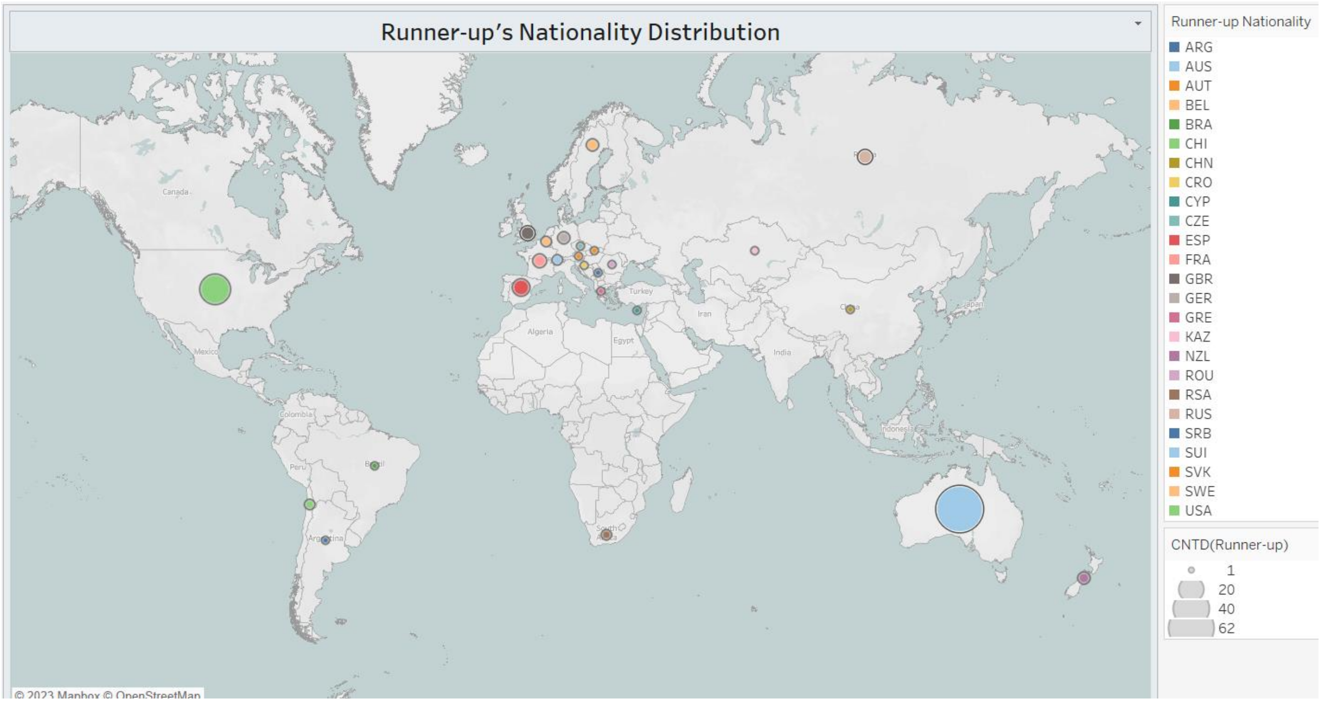

Figure 2: Runner-up's Nationality Distribution

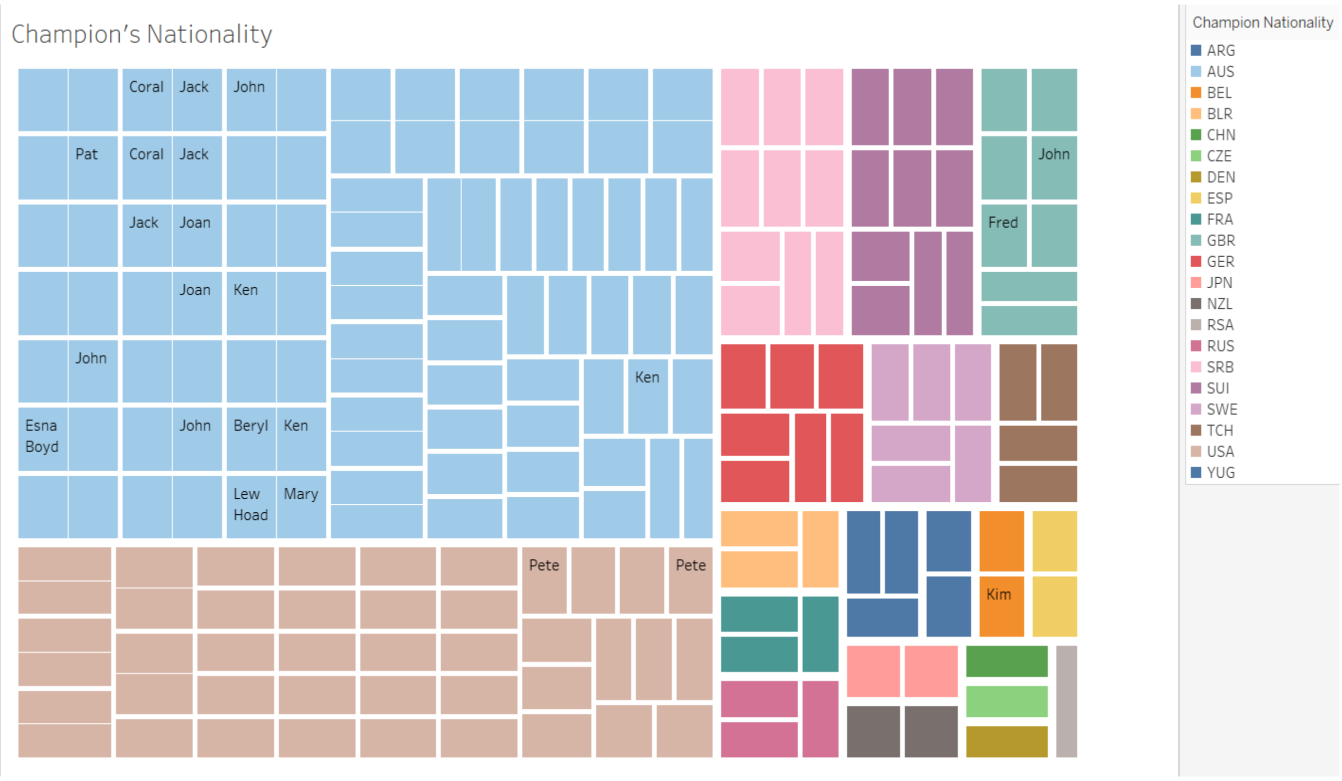

Figure 3: Champion's Nationality

Figure 1 and 2 shows the nationalities distribution of the champions and runners-up using a geographic map. The geographic map illustrates that Australia had the most count of people who are champion and runner-up, showcasing the fact that Australia is one of those countries that is actively participating with the open tennis championships. Figure 1 shows that Australia had the most champions over the past years with a distinct count of 44. Thus, the trend of Australia having most champions over other countries will be expected in the future. This can be further seen with the use of tree map in figure 3 where Australia had the most champions followed by USA. Figure 1 and 2 also illustrates that nationalities in Western Europe also participated actively in the tennis championships, along with its increasing number of champions, it is also expected that there will be an increasing growth of championships in the future.

Geographic map was used to effectively present the story and its context as the story focuses on nationalities. Champion and runner up's nationalities was labelled with different colours to help differentiate and along with the size of the circle that determines a distinct count of champions and runner-up in each nationality. Furthermore, for each nationality, the associated country names are included within the details of each country. This will overall improve the readability of the graph, hence improving its storytelling. Figure 3 have also illustrated number of champions across different nationalities using a tree map. Similarly, it can also display hierarchical representation of the context in an ease readability and storytelling format where users can directly see Australia with the most champions and following by USA.

Gender Trends

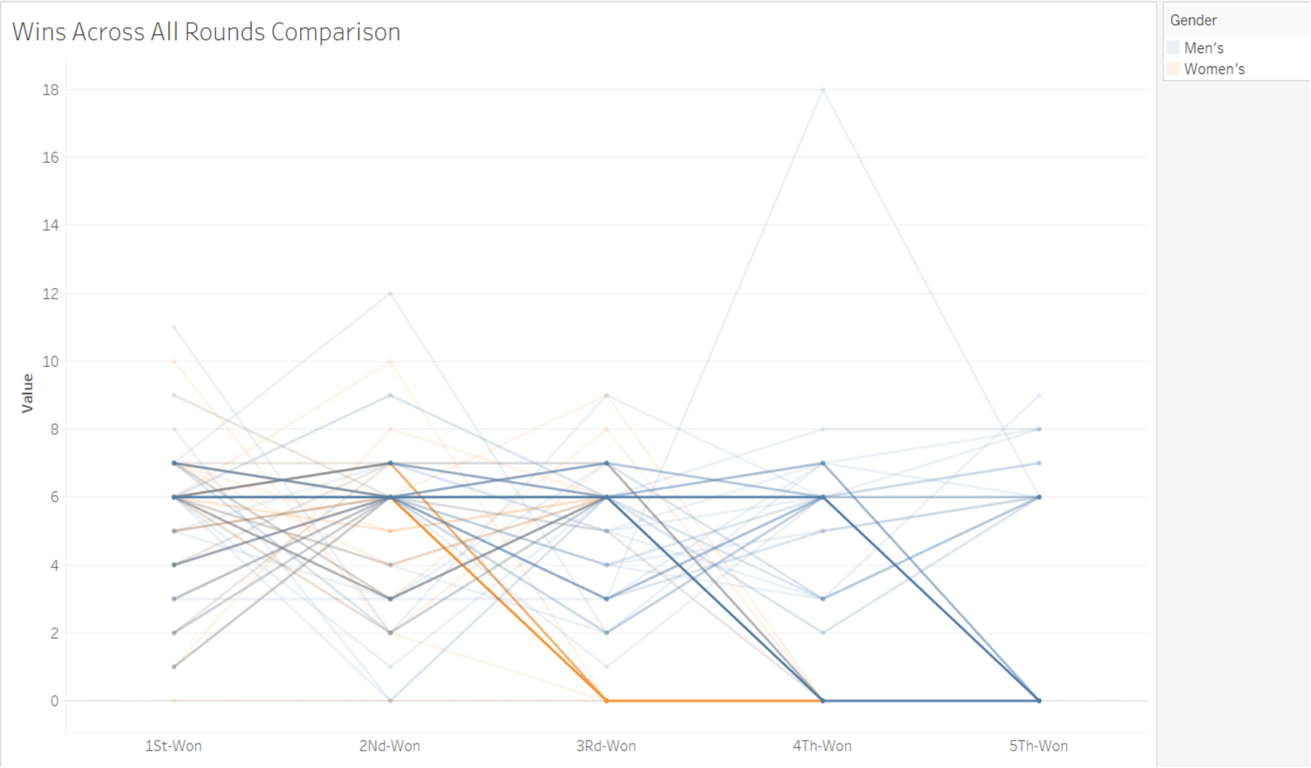

Figure 4: Wins Across All Rounds Comparison

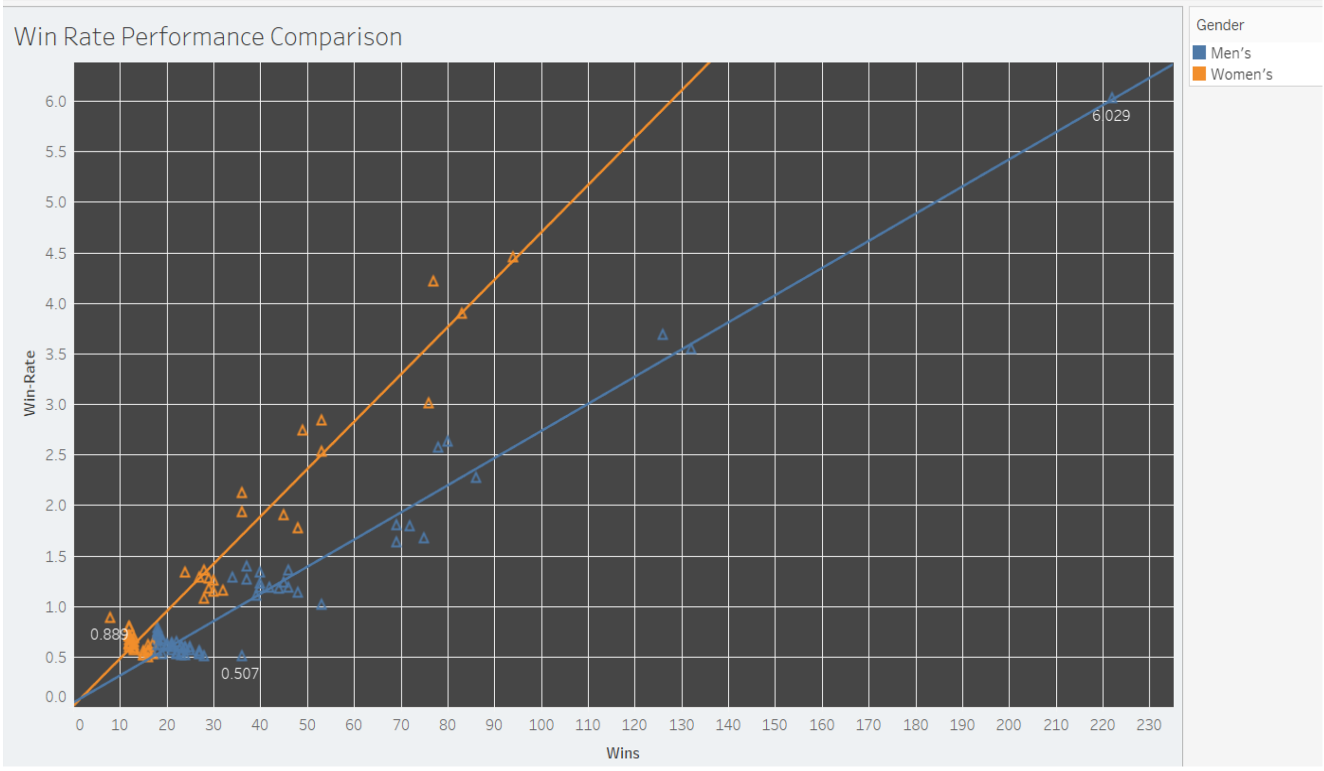

Figure 5: Win Rate Performance Comparison

The parallel coordinate in figure 4 illustrates wins in each round with the use of colour to differentiate between males and females. By decreasing opacity of the line, it can be clearly seen that most players have an average of 6 wins in each set of rounds while some go above 6 wins due to tie break. However, there is an outlier with a male player who had to win 18 games in order to win for a set round in comparison to others.

As such, the scatterplot in figure 5 shows that majority champions that only won one championship with less than 30 wins sits win rate of 1.0. One female champion had outperformed other champions with a win-rate of 88.9% in respective of her total wins of less than 10 while comparing to one the male champion who had 50.7% win-rate with almost 40 wins. The linear trending line between males and females shows as the total number of wins increases (increasing in number of championships), females tend to outperform males due to higher total of win rates with lower wins (lesser game played). There is also an outlier who had a highest total win rates and total wins which indicates he had won multiple championships in comparison to others.

The parallel coordinate was used to effectively illustrate the high dimensional data given by the large number of champions during analysis. These multidimensional data are represented in the form of line, it becomes easier to pick out trends and identify patterns. By using different colours, relationships between variables can be easily seen which enhances storytelling.

The scatterplot was used to compare two numerical variables where sum of wins is compared with the sum of win rates amongst champions to display its relationship. With the use of trending line, the trend between performances of males and females can also be effectively represented and along with labelling to ease readability, hence enhancing its storytelling.

CHANGES OVER TIME

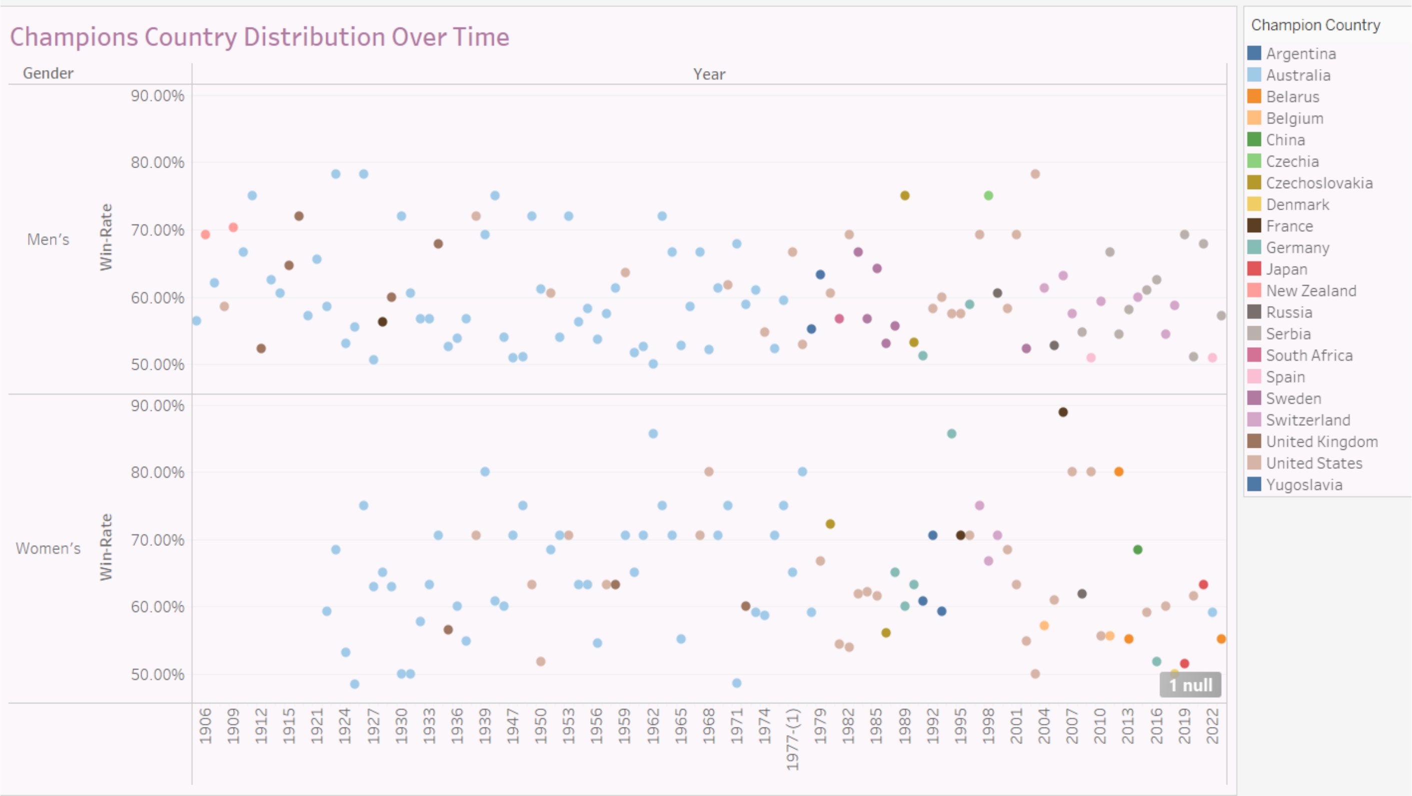

Figure 6: Champions Country Distribution Over Time

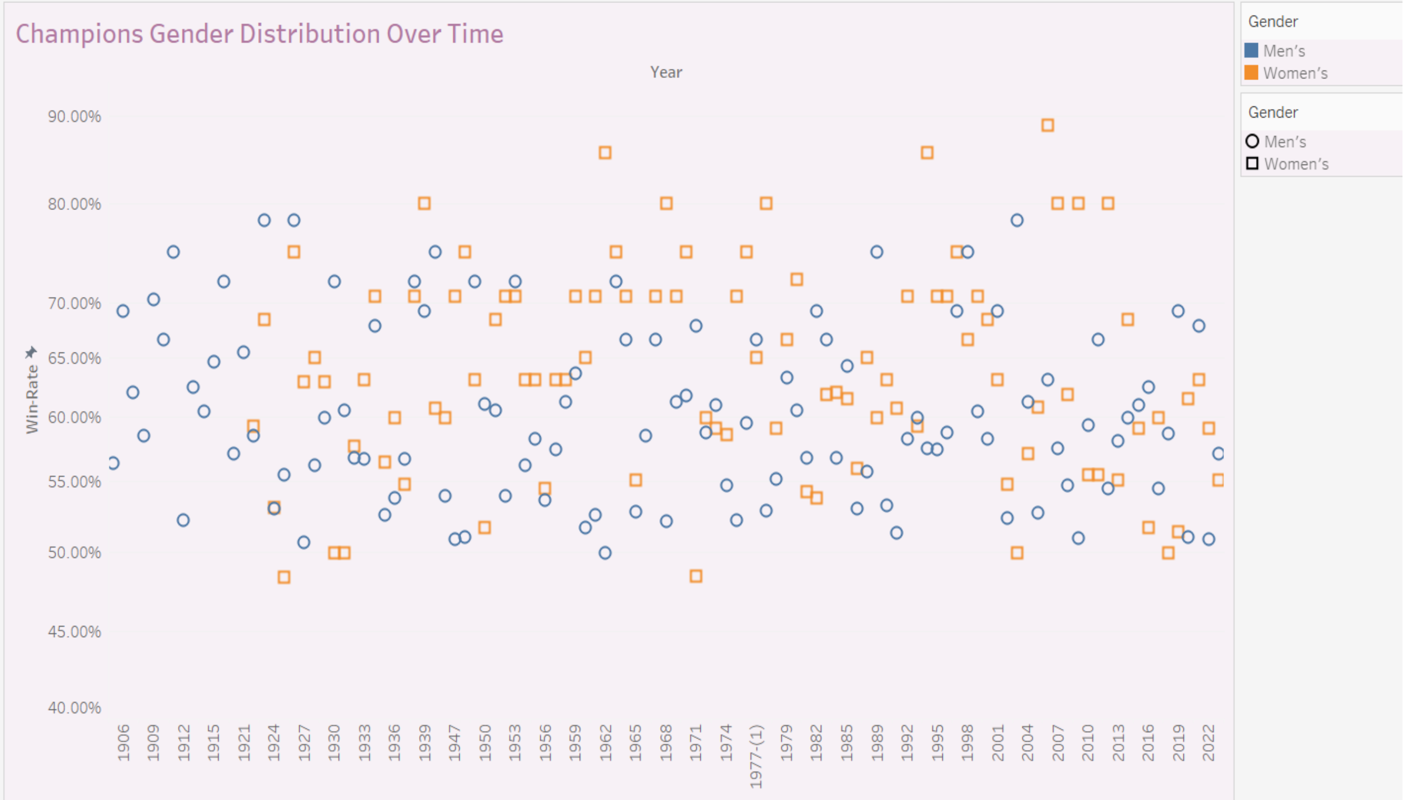

Figure 7: Champions Gender Distribution Over Time

Figure 6 shows the champions country distribution over time with different colours used to differentiate between the different countries. Each champion is split into males and females group in which it can be visibly seen that female participated and got the first championships later than males, after 1921. From figure 6, Australia was dominating the tennis championships from 1906 until late 1977. This was possibly due to tennis being an unknown sport for other countries and while technologies weren't advanced, other countries had limited ways of learning such sport. After 1977 as tennis becoming a more popular sport universally, more countries were able to learn and participate within the tennis championship, ultimately led to more champions within the countries. USA was one the countries that participated actively within those periods until now with a consistent increase in championships each year. Figure 7 further shows the gender distribution of the champions where females began having champions after 1921 and tend to achieve an average of higher win-rates than males, with some outlier champions who have achieved a win rate of more than 85%.

The use of scatter chart in both figure 6 and 7 provides a clear representation of changes over time by comparing the year and the win rates of champions, using different colours to convey different context and differentiate between its categories. This ultimately created a simple graph consisting with all data points, allows the users to easily see its trending as outlined with figure 6 where number of championships in Australia have been decreasing while countries such as USA have been increasing over time.

TOP PLAYERS PERFORMANCE VISUAL ANALYSIS

Win-Rate Performance

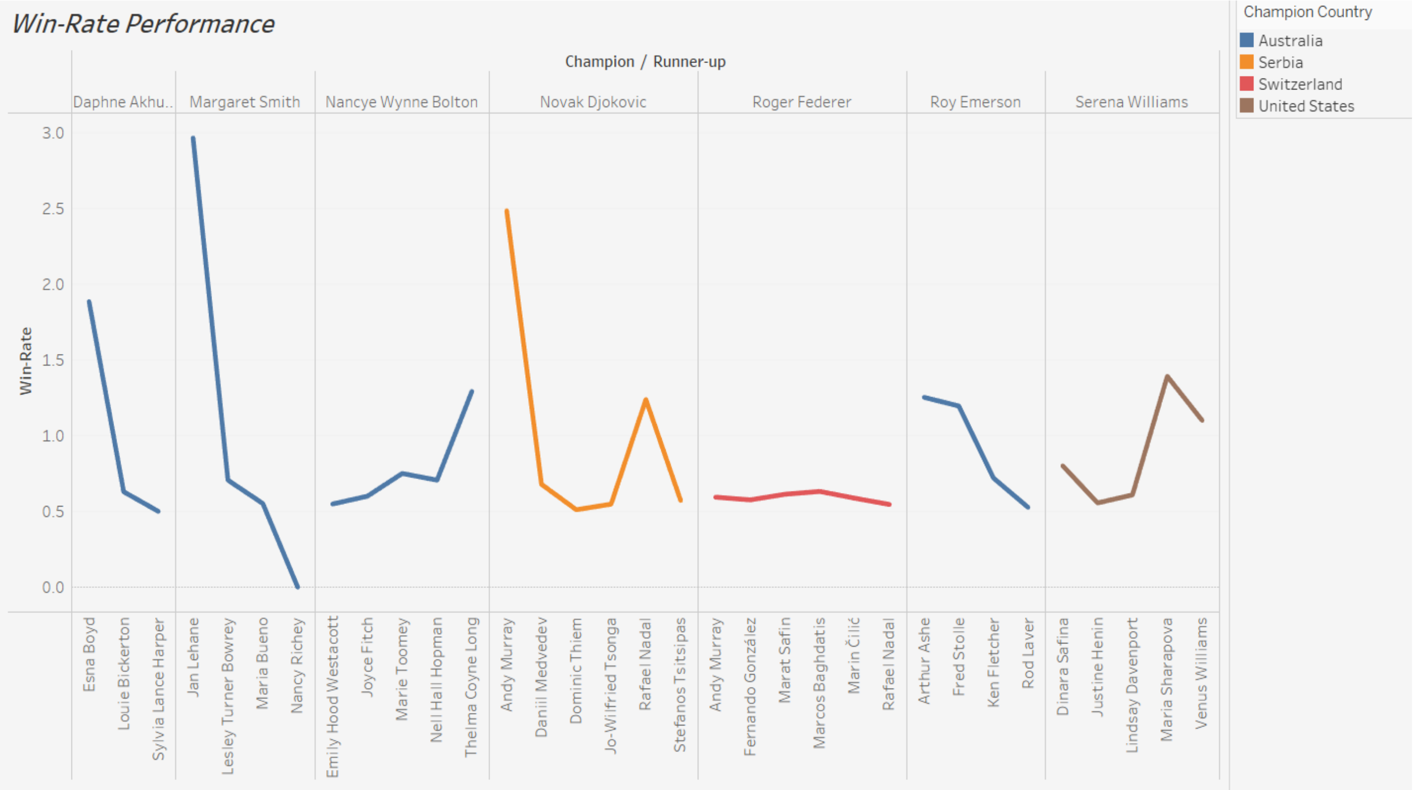

Figure 8: Top Players Win-Rate Performance

Figure 8 illustrates the top players win rate comparison along with their respective opponents. The graphic shows that most champions came from Australia with the use of colours to differentiate between countries. By contrasting the total win-rates of the top champions against each of their opponent, it can be clearly seen that Margaret had the highest total of win-rate against Jane, indicating that Margaret had against Jane multiple times and have won the championship within these games. The higher total win rates and the respective number of runners-up for each champion will indicate that champion had won more championships which can be seen with Margaret and Novak, both having a higher total of win rates than others, thus, indicating they have won the most championships in comparison to others.

By norm, each championship game will have maximum win rate of 1.0 (100%). The graph had shown an outlier with “Roger Federer” in comparison to other top champions where all his opponents were only played once, with an average win-rate of 50% for each game.

The benefit of using scatter chart, presenting the data points with lines gives an effective representation of comparison between champions total win-rates with their respective runner-up. It resulted the graphic with simplicity which will enhance its readability, allowing the user to understand and interpret the stories and context of the graph.

SET WIN-RATE PERFORMANCE

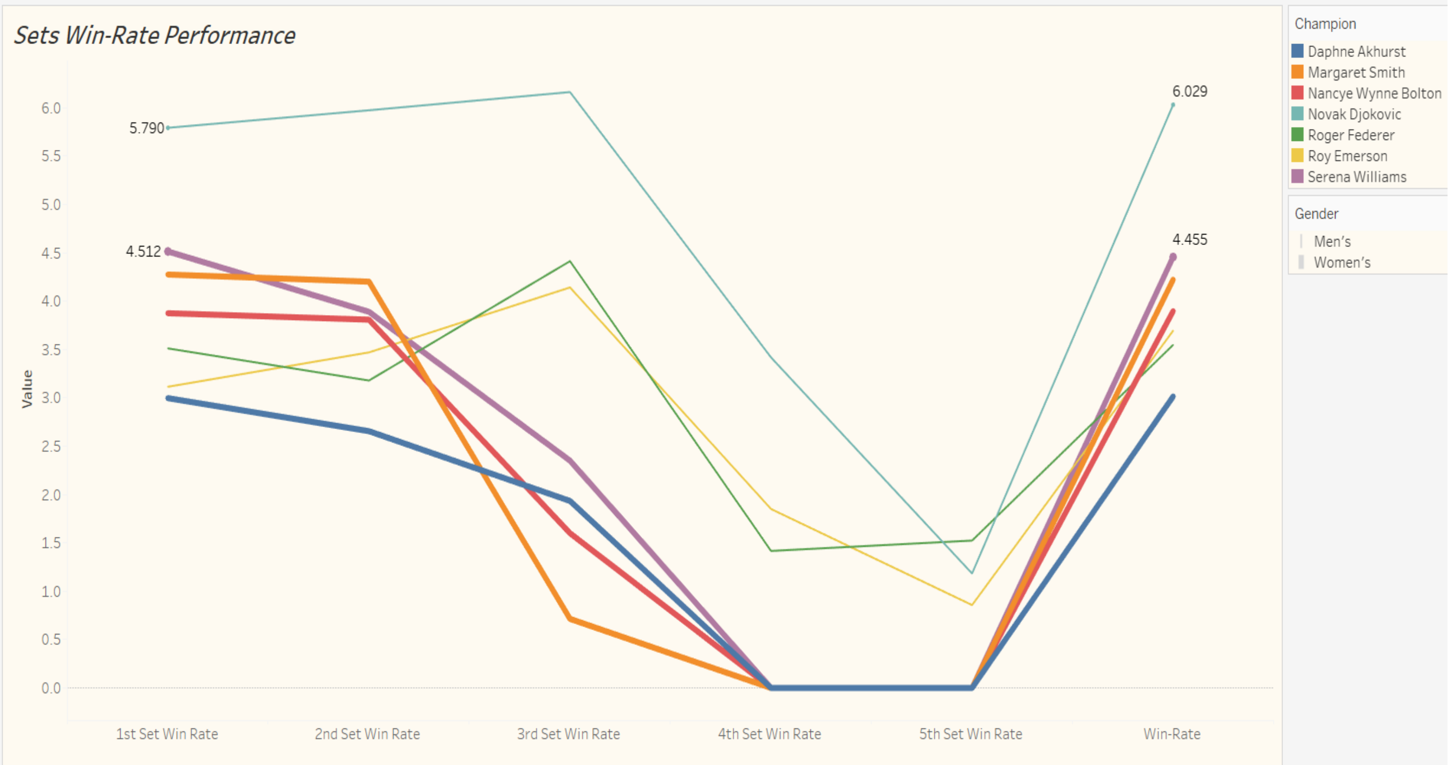

Figure 9: Top Players Sets Win-Rate Performance

Figure 9 illustrates the top players win rate performance across the 5 sets. Note that some champions had 0-win rate in 4th and 5th set as they were females which only required 3 set rounds to be played. As such. the gender of each champion has been differentiated by the size of its lines where thicker means they are female and thinner means they are males as indicated by the legends. Each champion is differentiated by using different colours.

By comparing the total sets win-rate across all champions, it can be clearly seen that Novak Djokovic had performed better than other champions in almost all sets round, with the highest win-rate in set 1 of 5.790, resulting with the highest win rate of 6.029. By gender, Novak outperformed top champions who were males while Serena outperformed top champions who were females with the win-rate of 4.455. Out of all champions, Daphne performed the lowest with its win-rate across set 1,2 and 3, resulting with the lowest win-rate.

The parallel coordinate in figure 9 was used to clearly present the set win-rates across different champions, providing the chart with simple and informative representation of the context. This allows the user to easily compare the win-rates performance across different top champions which showcasing the purpose and effectiveness of the graph in storytelling. The use of labels in top champion for each gender group further allows the user to see which champion performed the best within their gender group.

ADVANTAGES AND DISADVANTAGES OF TABLEAU AND ITS VISUALIZATION TECHNIQUES

Advantages	**Disadvantages**
Ease of use	Lack of direct editing of data
Real time data updates	Limited customization/visualizations – users are only limited to the graphs that tableau has
Quick prototyping	high cost for licenses
Ability to handle large dataset	Potential misinterpretation of data
Able to create various distinct visualisation graphs such as tree maps, parallel coordinates and geographic map which supports high dimensional data analysis	Lack of advanced statistical analysis tools
The use of colours, size, details, labels to differentiate and convey important information – enhance storytelling	Needed manual effort to handle the graph for people with special need (e.g., colour blind)
Able to display and compare between multiple variables, both in columns and rows	No automatic saving of the files – may result in loss of progress

Ultimately, Tableau is easy to navigate (user friendly) and able to handle large datasets which can be analysed and presented in various visualization techniques such as parallel coordinate where other tools such as Excel, won't have these types of complex visualizations. Also, a great tool for experimenting between different variables to deep into different insights. However, it is expensive to be able to use Tableau and any changes to the dataset needed to be changed within other platform such as Excel to be able to update it and use it for visualizations in Tableau.

CONCLUSION

To conclude, countries that had the most champions is Australia regarding nationalities trends displayed in figure 1, 2 and 3. However, through analysing changes over time, the number of champions in Australia have been decreased dramatically after 1977 while other countries such as USA slowly increasing, showcasing the trend that USA may become the top country with most championships won. The gender trend analysis discovered that females tend to have higher win-rates than males, in respective to same number of games played. Finally, the performance chart of the top champions who have won five games or more shows that 4 out of 7 champions were from Australia and out of the 7 champions, Novak Djokovic outperformed others in terms of the number of championship Novak have won and its win-rate being the highest overall.